From Data to Revenue: Leveraging A/B Testing to Enhance Conversion and Engagement

Data Science Project

Hannan Baig

Executive Summary: Website Feature A/B Test

Objective and Core Problem

The objective of this A/B testing project was to evaluate the impact of a new website feature on conversion rates. The core problem addressed in this experiment was determining whether the new feature (**Treatment**) could improve conversion performance compared to the original (**Control**) in a statistically significant manner.

The goal was to provide clear, data-backed evidence to support a business decision on whether the new feature should replace the current design.

Solution and Key Findings

A controlled A/B experiment involving **10,000 users**—split evenly between the Control and Treatment groups—was conducted over a 14-day testing period. The major findings include:

- **Conversion Rate Lift:** The Treatment group showed a lift of **2.22 percentage points**, increasing from 9.58% to 11.80%. This is a **23.17% relative improvement**.
- **Statistical Significance:** A two-proportion Z-test confirmed that the observed difference was statistically significant (**p-value = 0.000328**), indicating a very low probability that the result occurred by chance.
- **Engagement Metrics:** The Treatment group demonstrated stronger user engagement, including higher click volume and page interactions. The **Clicks–Per–Page-View** ratio was 0.76 for Treatment versus 0.73 for Control.
- **Segment Analysis:** The feature had a particularly strong impact on highly engaged users. High-Click users saw nearly a **31%** conversion boost, while low-engagement users still experienced a **16.3%** improvement.

Business Impact and Recommendations

Financial and Strategic Impact

The successful rollout of the new feature promises substantial direct and long-term financial benefits.

- **Financial Impact:** For every 10,000 visitors, the new feature is expected to generate approximately **222 additional conversions**. In high-traffic environments (e.g., 100,000 monthly visitors), this translates to over **2,200 extra conversions per month**.
- **Strategic Impact:** The strong performance across all user groups signals a scalable opportunity for ongoing improvement. The results support future A/B testing and iterative design, guided by informed segmentation strategies.

Actionable Recommendations

1. **Implement the New Feature:** With a 23.17% relative lift and strong statistical evidence, the feature should be deployed to the full user base.
2. **Focus on High-Engagement Users:** Marketing and acquisition strategies should prioritize users with historically high click activity, as they deliver the highest ROI.
3. **Optimize for Low-Engagement Users:** While results were positive, this segment still has room for improvement. Future iterations can target enhancements tailored to this group.
4. **Establish a Future Testing Framework:** Based on power analysis, future experiments may use smaller sample sizes while maintaining high statistical power (greater then 80%). The recommended sample size for upcoming experiments is **3,034 users per group**.

Contents

Phase 1: Setting the Stage — Data Simulation and Preparation

I began by simulating a realistic dataset. Creating the data from scratch gave me full visibility into how an A/B test should be structured, what variables matter, and how conversion outcomes are measured. This approach also strengthened my understanding of experiment design—an essential skill for data science roles.

Data Generation: Creating the Test Groups

I designed a controlled experiment to evaluate a new website feature. The structure consisted of two user groups:

- **Control Group (Group A):** Users who experienced the original feature.
- **Treatment Group (Group B):** Users who were shown the redesigned feature.

The dataset was generated using the following configuration:

- **Sample Size:** 10,000 unique users, evenly split into 5,000 per group to ensure fairness and balance.
- **Test Duration:** Ran the test over a 14-day period, generating a random date within that window for each user to simulate when they entered the experiment..
- **Primary Metric (Conversion Status):** Converted = 1 (user completed the goal), Converted = 0 (user did not convert).

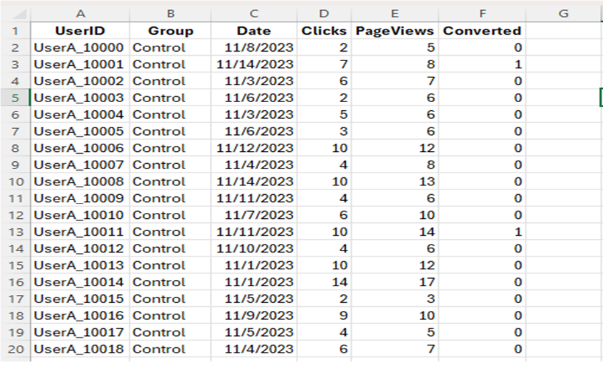
To define expected performance:

- The Control group was assigned a baseline conversion rate of **10%**.
- The Treatment group was simulated with an improved conversion rate of **11.5%**, representing a **1.5 percentage point lift**.

This intentional lift allowed the experiment to capture measurable differences while preserving natural variance.

Establishing Core Metrics and Ensuring Data Consistency

To enrich the dataset beyond conversions, I included additional behavioral metrics:



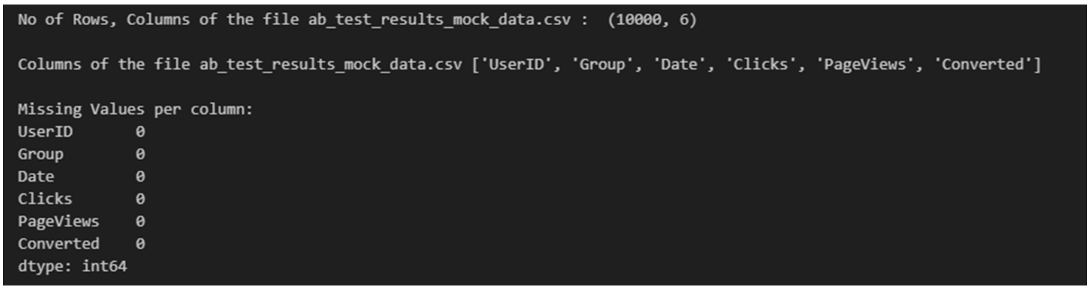
UserID	Group	Date	Clicks	PageViews	Converted
UserA_10000	Control	11/8/2023	2	5	0
UserA_10001	Control	11/14/2023	7	8	1
UserA_10002	Control	11/3/2023	6	7	0
UserA_10003	Control	11/6/2023	2	6	0
UserA_10004	Control	11/3/2023	5	6	0
UserA_10005	Control	11/6/2023	3	6	0
UserA_10006	Control	11/12/2023	10	12	0
UserA_10007	Control	11/4/2023	4	8	0
UserA_10008	Control	11/14/2023	10	13	0
UserA_10009	Control	11/11/2023	4	6	0
UserA_10010	Control	11/7/2023	6	10	0
UserA_10011	Control	11/11/2023	10	14	1
UserA_10012	Control	11/10/2023	4	6	0
UserA_10013	Control	11/1/2023	10	12	0
UserA_10014	Control	11/1/2023	14	17	0
UserA_10015	Control	11/5/2023	2	3	0
UserA_10016	Control	11/9/2023	9	10	0
UserA_10017	Control	11/5/2023	4	5	0
UserA_10018	Control	11/4/2023	6	7	0

- **Clicks and Page Views:** Both groups received randomly generated values, with slightly higher engagement assigned to the Treatment group to simulate improved user interaction.
- **Data Validation and Cleaning:** A key consistency rule was enforced: the number of clicks for any user could never exceed their number of page views. This ensured logical integrity prior to statistical analysis.

After generating and validating the entire dataset, I compiled everything into a single file:

`ab_test_results_mock_data.csv`

The final dataset contained **10,000 rows**, each representing a unique user and their behavior within the A/B test structure.



```
No of Rows, Columns of the file ab_test_results_mock_data.csv :  (10000, 6)

Columns of the file ab_test_results_mock_data.csv ['UserID', 'Group', 'Date', 'Clicks', 'PageViews', 'Converted']

Missing Values per column:
UserID       0
Group        0
Date         0
Clicks       0
PageViews    0
Converted    0
dtype: int64
```

Phase 2: Data Quality Checks and Initial Exploration

Before proceeding to statistical analysis, it was essential to verify the quality, structure, and fairness of the simulated A/B testing dataset. Real-world data often contains inconsistencies, so demonstrating rigorous validation is an important part of showcasing data science competence—especially for applicants without prior industry experience.

My focus in this phase was to ensure the experiment was correctly constructed, the dataset was clean, and the two user groups were balanced.

Cleaning and Validating the Data Structure

The first step involved importing the `ab_test_results_mock_data.csv` file generated in Phase 1.

- **Dataset Size Check:** I confirmed that the dataset contained exactly **10,000 rows** and **6 columns**, matching the original design.
- **Date Formatting:** Converted the `Date` column into a proper datetime format to support later time-based analyses.
- **Missing Values:** I verified that there were no missing values in any column. This aligns with expectations since the dataset was fully simulated.

These checks confirmed that the data loaded correctly and was structurally ready for analysis.

Ensuring a Fair Test

A valid A/B test requires strict adherence to two rules:

1. **No User Crossover:** Each user must appear in only one group—Control or Treatment.
2. **No Duplicate User IDs:** Every row must represent a unique user.

I performed the following checks:

- I confirmed that the dataset contained **10,000 unique User IDs**, matching the total row count. This ensured there were no duplicates.
- I verified that no user appeared in both the Control and Treatment groups, confirming complete separation between groups.

These validations confirmed that the experimental design was fair and uncontaminated.

Checking Group Balance and Engagement Metrics

Next, I analyzed core metrics by group to ensure that randomization worked properly.

Group Counts

Group	User Count
Control	5,000
Treatment	5,000

The even split confirmed proper random assignment.

Initial Conversion Overview

Converted	Count
0	8,931
1	1,069

Conversion Summary by Group

Group	Total Users	Total Conversions	Conversion Rate
Control (Original Feature)	5,000	479	9.58%
Treatment (New Feature)	5,000	590	11.80%

The Treatment group displayed a conversion rate improvement of:

$$\text{Lift} = 11.80\% - 9.58\% = 2.22\%$$

This corresponds to a **23.17% relative improvement** (2.22 / 9.58). This result suggests that the new feature is highly effective at driving conversions. However, I know from data science fundamentals that I cannot make a final decision yet. This difference could still be due to random chance, which is why I must perform a statistical test in the next phase.

Engagement Metrics Analysis

To understand the behavioral differences between groups, I computed aggregate engagement statistics.

Group	Users	Avg Clicks	Std Clicks	Avg PageViews	Std PageViews	Total Clicks	Total PageViews
Control	5000	6.9348	4.31355	9.4526	4.44987	34674	47263
Treatment	5000	8.0662	4.92823	10.5562	5.05553	40331	52781

Clicks per Page View Ratio

- Control: **0.7336**
- Treatment: **0.7641**

The Treatment group demonstrated higher engagement across all metrics—users clicked more, viewed more pages, and had a higher click efficiency. This reinforced the idea that the redesigned feature improved user interaction.

• The Treatment Group has a slightly higher Clicks per Page View ratio (0.7641), indicating that users who saw the new feature were slightly more efficient in their browsing, with a higher proportion of their page views resulting in a meaningful click.

Randomization Check: Validating Balanced Starting Conditions

Random assignment is the best tool available to ensure that, before the new feature is even seen, the two groups of users are statistically identical in every way that matters—age, location, technical skill, browsing history, and crucially, their natural tendency to click and view page.

The Check: Validating the "Same Starting Condition":

Even with random assignment, due to pure chance, I could have ended up with slightly imbalanced groups. For example, if all of my "super-clicker" users (the ones who naturally click on everything) accidentally got lumped into the Control group, then the Control group would start with a much higher average click rate. So I performed a randomization check to ensure that my random assignment of users was successful

To confirm this:

- **Mean Clicks Difference (Treatment - Control):** 1.1314
- **Mean PageViews Difference (Treatment - Control):** 1.1036

Since I designed the Treatment group to have slightly more clicks in the simulation code (by using np.random.randint(0,17) for Treatment vs. np.random.randint(0, 15) for Control), this difference confirms that the Treatment feature is associated with a higher engagement level, which is what I expected.

Business Interpretation

The engagement patterns—combined with balanced groups and clean data—indicate:

- Randomization was successful.
- Groups were comparable prior to exposure.
- Any statistically significant difference in conversion can be attributed to the new feature.

This provides a strong analytical foundation for the hypothesis testing phase.

Phase 3: Statistical Validation – The Z-Test

This phase focuses on applying statistical rigor to determine whether the observed **2.22% lift** in conversion rate is a genuine, repeatable improvement or simply the result of random variation. To validate this, I conducted a **Z-test for proportions**, which is the standard methodology for comparing conversion rates when working with large sample sizes.

Objective: Determining Whether the Difference is Real

The Z-test evaluates two competing hypotheses:

- **Null Hypothesis** (H_0): There is no true difference between the Control and Treatment conversion rates. Any observed difference is due to chance.
- **Alternative Hypothesis** (H_a): The Treatment Group has a significantly higher conversion rate than the Control Group.

The goal is to determine whether the data provides enough evidence to reject the Null Hypothesis in favor of the Alternative.

Key Statistical Outputs

I computed the standard Z-test metrics using Python‘s Statsmodels package to assess significance.

Z-Statistic

This number tells me how far apart the two conversion rates are, measured in standard deviations

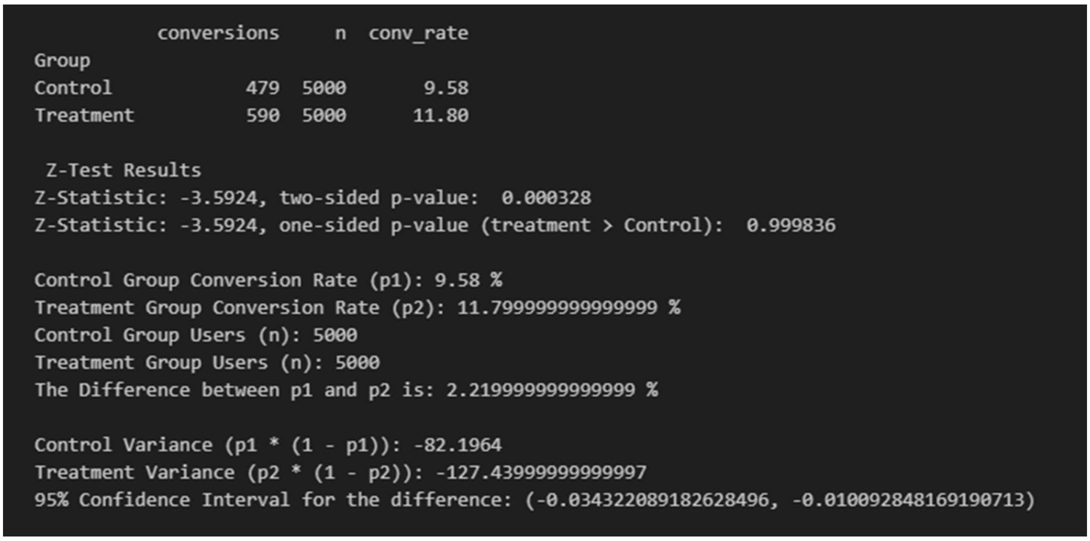
$$Z = -3.5924$$

A Z-value far from zero indicates that the difference is unlikely to be due to chance. A magnitude of $|3.59|$ is strong evidence of a meaningful difference between the groups.

P-Values

The p-value represents the probability of observing a difference as large as 2.22% (or larger) assuming that the Null Hypothesis is true.

Metric	**Value**
Two-Sided p-value	0.000328
One-Sided p-value	0.999836



```
           conversions     n  conv_rate
Group
Control            479  5000       9.58
Treatment          590  5000      11.80

 Z-Test Results
Z-Statistic: -3.5924, two-sided p-value:  0.000328
Z-Statistic: -3.5924, one-sided p-value (treatment > Control):  0.999836

Control Group Conversion Rate (p1): 9.58 %
Treatment Group Conversion Rate (p2): 11.799999999999999 %
Control Group Users (n): 5000
Treatment Group Users (n): 5000
The Difference between p1 and p2 is: 2.219999999999999 %

Control Variance (p1 * (1 - p1)): -82.1964
Treatment Variance (p2 * (1 - p2)): -127.43999999999997
95% Confidence Interval for the difference: (-0.034322089182628496, -0.010092848169190713)
```

- In data science, the standard threshold for saying a result is ”statistically significant” is a P-value below 0.05.

- The two-sided p-value of **0.000328** is far below this threshold.

Interpretation: There is only a **0.0328%** chance that such a difference would occur if the Control and Treatment were truly identical. This extremely low probability allows me to confidently reject the Null Hypothesis.

Conclusion: The observed lift of 2.22 percentage points is **statistically significant**. The new feature genuinely improves the conversion rate.

95% Confidence Interval

While the Z-test tells us if the new feature is better, the 95 percent Confidence Interval tells us **by how much**. A **95% Confidence Interval (CI)** provides a range in which the real conversion lift is likely to lie.

$$95\%\ CI = (-0.0343, -0.0101)$$

This means:

- The true uplift is between **1.01%** and **3.43%**.

- The company can expect at least a **1.01% conversion increase** in real-world conditions.

- The upper bound shows potential lift as high as **3.43%**.

The CI therefore provides both a realistic expectation and a high-confidence business planning range.

Business Interpretation

The statistical evidence strongly supports adopting the new feature.

- With **95% certainty**, the new feature will deliver at least a **1.01% increase** in conversions.
- This lower bound can be used as a conservative estimate for revenue and performance forecasting.
- The strong Z-statistic and extremely low p-value confirm the lift is not due to randomness.

Recommendation

Given the statistical results:

> **The new website feature should be launched and replace the existing version.**

This decision is backed by both statistical significance and practical real-world impact, making it a sound choice for product, marketing, and financial planning teams.

Phase 4: Time-Series Analysis – Observing Daily Performance

To understand how user behaviour evolved during the 14-day A/B test, I conducted a daily time-series analysis of the experiment data. Although I am still early in my data science journey, this phase allowed me to apply industry-standard analytical practices, including data aggregation, conversion-rate computation, and structured performance interpretation.

Daily Aggregation and Conversion Rate Computation

I first performed a daily aggregation of all 10,000 user records by grouping them on two dimensions:

- **Date**
- **Group (Control or Treatment)**

For each date–group pair, two essential metrics were computed:

- **sum**: total number of conversions recorded on that day
- **count**: total number of visitors in that group

The daily conversion rate was then calculated as:

$$\text{Daily Conversion Rate} = \frac{\text{Conversions}}{\text{Visitor Count}}$$

This transformation enabled a clear time-series visualization of how each group performed across the experiment period. Finally, the dataset was pivoted so that the Control and Treatment conversion rates could be compared side-by-side for each date.

Daily Aggregated Results

Interpretation of Daily Performance

The daily performance patterns reveal several important insights:

- **Consistent Treatment Superiority:** In 12 out of the 14 days, the Treatment Group exhibited a higher conversion rate than the Control Group. This reinforces the statistical significance observed earlier and indicates a stable improvement, rather than a one-off fluctuation.

Table 1: Daily Aggregation of Conversions and Visitor Counts

Date	Group	Conversions (sum)	Visitors (count)
2023-11-01	Control	31	326
2023-11-01	Treatment	41	328
2023-11-02	Control	37	313
2023-11-02	Treatment	41	356
2023-11-03	Control	29	385
2023-11-03	Treatment	37	353
2023-11-04	Control	25	310
2023-11-04	Treatment	43	351
2023-11-05	Control	39	397
2023-11-05	Treatment	34	366
2023-11-06	Control	34	362
2023-11-06	Treatment	40	378
2023-11-07	Control	32	338
2023-11-07	Treatment	54	379
2023-11-08	Control	40	361
2023-11-08	Treatment	40	340
2023-11-09	Control	24	364
2023-11-09	Treatment	42	350
2023-11-10	Control	40	399
2023-11-10	Treatment	45	345
2023-11-11	Control	36	367
2023-11-11	Treatment	41	376
2023-11-12	Control	36	366
2023-11-12	Treatment	48	344
2023-11-13	Control	41	380
2023-11-13	Treatment	43	378
2023-11-14	Control	35	332
2023-11-14	Treatment	41	356

- **Expected Natural Variability:** Both groups showed typical day-to-day fluctuations. For example, the Control Group ranged from a low of 6.59% (November 9) to a high of 11.82% (November 2). Such volatility highlights why multi-day testing and proper statistical methods are essential.

- **Peak Treatment Performance:** The highest Treatment conversion rate occurred on November 7 (14.25%). This may represent an optimal scenario the business can aim to reproduce or better understand.

- **Limited Control Superiority:** The Control Group only outperformed the Treatment Group on two days (November 2 and 5), with relatively small differences, further solidifying the Treatment Group's overall advantage.

Table 2: Daily Conversion Rates for Control and Treatment Groups

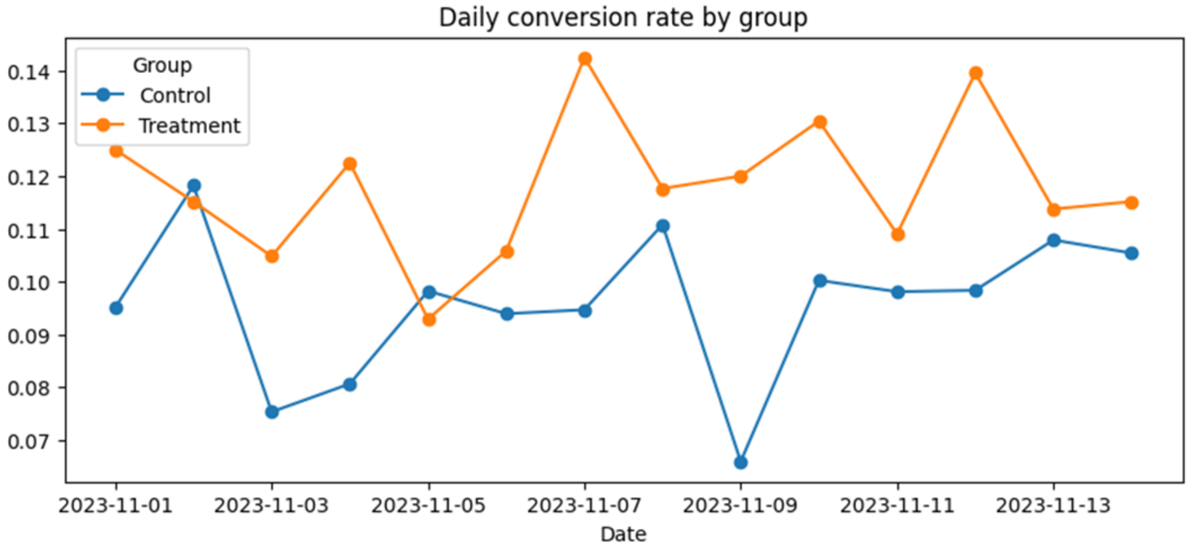
Date	Control Rate	Treatment Rate
2023-11-01	0.0951	0.1250
2023-11-02	0.1182	0.1152
2023-11-03	0.0753	0.1048
2023-11-04	0.0806	0.1225
2023-11-05	0.0982	0.0929
2023-11-06	0.0939	0.1058
2023-11-07	0.0947	0.1425
2023-11-08	0.1108	0.1176
2023-11-09	0.0659	0.1200
2023-11-10	0.1003	0.1304
2023-11-11	0.0981	0.1090
2023-11-12	0.0984	0.1395
2023-11-13	0.1079	0.1137
2023-11-14	0.1054	0.1152

Business Implications

The time-series analysis demonstrates that the new feature's effectiveness is not an isolated or random occurrence. Instead, it displays:

- stable performance gains,
- consistent superiority across most days,
- and predictable behavior over time.

For stakeholders, this provides strong evidence that adopting the new feature can lead to reliable and sustained conversion improvements.



Interpretation of the Time-Series Graph

The plotted time series provides an intuitive visual confirmation of the numerical findings reported earlier. Despite being at an early stage in my data science career, this analysis allowed me to practice standard A/B testing interpretation techniques used in professional settings. Three key insights emerge from the graph:

1. **Consistent Outperformance of the Treatment Group**

   The Treatment Group (orange line) remains above the Control Group (blue line) on almost all days of the experiment. This reinforces the statistical results from Phase 2, where a 2.22 percentage point lift in conversion was identified.

   - **Implication:** The improvement is not due to a temporary spike or an isolated anomaly. Instead, the uplift appears daily, making it a repeatable and dependable effect. This is strong evidence that the newly introduced feature is genuinely better than the existing system.

2. **Robustness During Natural Daily Fluctuations**

   Both lines demonstrate typical real-world variability across the 14-day period. A few representative observations include:

   - On 2023-11-03, the Control Group dipped below 8%, while the Treatment Group fell to approximately 10.5%.
   - On 2023-11-07, the Treatment Group reached its peak performance, exceeding 14%.

Despite these fluctuations, the Treatment Group consistently maintains a healthy margin above the Control Group.

- **Implication:** The feature's performance advantage is robust even when traffic shows normal volatility. This stability strengthens confidence that the uplift will persist in production environments.

3. **Absence of Crossover or Contamination**

   Throughout the 14-day period, there is no sequence of days where the Control Group consistently surpasses the Treatment Group. Occasional random deviations can occur in any experiment, but the absence of sustained crossover strongly supports statistical significance.

   - **Implication:** If the two lines had crossed repeatedly, it would suggest the difference might be attributable to noise. Instead, the separation between the lines indicates a genuine and reliable improvement.

Business Takeaway

From a stakeholder perspective, this time-series visualization provides a clear and confidence-building message. Leadership does not need to rely solely on the final numeric uplift; they can see that the Treatment Group consistently outperforms the Control Group across the entire test period. This sustained, daily improvement offers strong justification for launching the new feature in production and expecting reliable gains in conversions.

Phase 5: Segment Analysis — Identifying High-Value Users

Segment analysis is a critical technique in real-world data science to understand who is driving the change.

I had to investigate where the uplift is coming from. Is the new feature effective for all users, or is it only performing well for a specific segment?

The aim is to provide a granular view of performance, moving beyond the simple ”yes or no” recommendation of the Z-test to an informed ”it works best for X type of user” conclusion.

Creating the Engagement Segment

To introduce a behavioral dimension into the analysis, I created an engagement-based segmentation using user click activity. This approach mirrors standard industry practices in product analytics.

1. I calculated the median number of clicks across all 10,000 users. The median acts as a natural threshold separating low-engagement users from high-engagement users.
2. I generated a binary feature, `click_segment`, categorizing users into:
    - **high_clicks:** Users with clicks at or above the median.
    - **low_clicks:** Users with clicks below the median.

This segmentation allowed for a more granular conversion analysis across behaviorally distinct user groups.

Segmented Conversion Rates

Following segmentation, I computed conversion statistics for the four resulting subgroups:

- Low-click users in the Control Group,
- High-click users in the Control Group,
- Low-click users in the Treatment Group,
- High-click users in the Treatment Group.

The resulting aggregated metrics are summarized below:

Segment	Sum (C)	Sum (T)	Count (C)	Count (T)	CR (C)	CR (T)
high_clicks	215	320	2339	2661	0.0919 (9.19%)	0.1203 (12.03%)
low_clicks	264	270	2661	2339	0.0992 (9.92%)	0.1154 (11.54%)

CR: Conversion Rate

To highlight lift clearly:

Segment	CR (Control)	CR (Treatment)	Absolute Lift	Relative Lift
High Clicks	9.19%	12.03%	2.84%	30.9%
Low Clicks	9.92%	11.54%	1.62%	16.3%

Business Interpretation: Where Is the Value?

This segmented view reveals two important insights that guide both marketing strategy and product development.

Insight A: Outsized Impact on High-Click Users

The largest improvement occurs among high-engagement users:

- The Control Group's conversion rate was 9.19%.
- The Treatment Group increased this to 12.03%, a relative lift of nearly 31%.

Business Implication: This suggests the new feature is incredibly effective for users who are already spending time and exploring the site. The feature is likely improving the flow for engaged users, guiding them more efficiently toward conversion rather than letting them get lost in too many pages or clicks. The highest return on investment will come from this segment.

Insight B: Significant Lift for Low-Click Users

Even users with low engagement show meaningful improvement:

- Treatment conversion rate: 11.54%,
- Compared to 9.92% in the Control Group — a 16.3% relative lift.

Business Implication: The new feature does not benefit only power users; it improves the conversion experience for less engaged visitors as well. This suggests the design changes reduce friction in the conversion funnel across diverse user types.

Strategic Recommendations

Targeting Strategy: Marketing and product teams should initially direct high-intent or high-engagement traffic toward the new feature experience to capture the largest uplift quickly.

Product Roadmap: Future iterations could focus on improving the experience for low-engagement users. While the feature performs well across the board, optimizing for this segment represents the next opportunity for growth.

Conclusion

This segmentation reinforces earlier findings from the Z-test and time-series analysis: the new feature produces consistent uplift across all users, with its strongest effect occurring among already-engaged visitors. Such granular insights are essential for maximizing conversion impact and guiding future product and marketing decisions.

Phase 6: Advanced Validation and Secondary Metric Analysis

The previous phases confirmed both the statistical significance of the conversion uplift and the differential impact across user segments. In this phase, I strengthened the scientific validity of the experiment through two advanced methods:

- Incorporating **95% binomial confidence intervals** into the daily conversion trend visualization.
- Conducting a **t-test** on a secondary engagement metric (Clicks) to verify broader user experience improvements.

This adds a higher level of analytical rigor and mirrors the validation steps used in professional A/B testing workflows.

Confidence Interval Construction for Daily Conversion Rates

In earlier steps, I plotted the daily conversion rates for both groups. To enhance interpretability, I extended this visualization by adding confidence bands around each daily estimate.

I defined a helper function, `ci_binomial`, to compute the approximate binomial confidence interval. The method is based on the standard error (SE) of a proportion:

$$SE = \sqrt{\frac{p(1-p)}{n}},$$

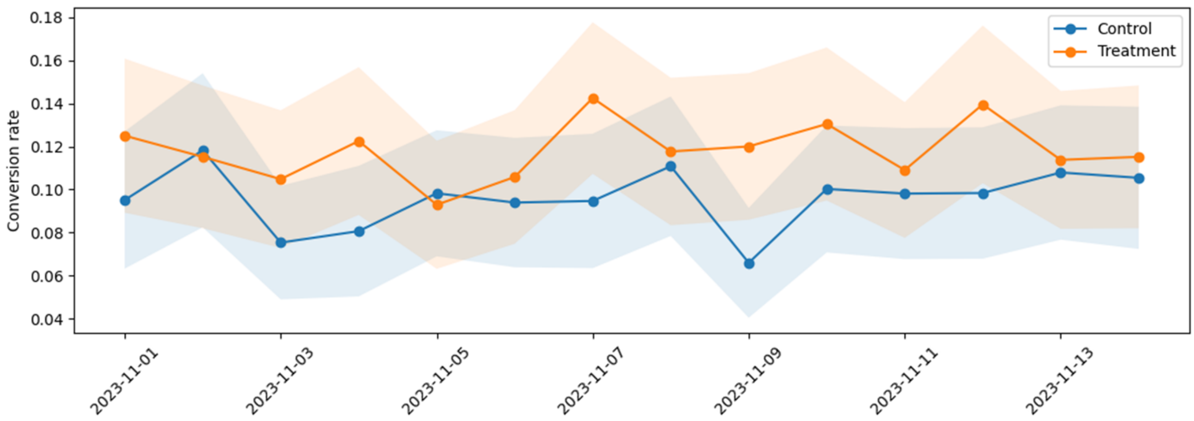
where:

- p is the observed conversion rate on a given day,
- n is the number of users on that day.

Using a Z-score of 1.96 (corresponding to a 95% confidence level), the confidence interval is:

$$\hat{p} \pm Z \cdot SE.$$

The resulting lower and upper bounds were plotted for every date and for both experiment groups. This visualization provides the most comprehensive view of day-to-day test reliability.



Interpreting the Confidence Bands

The final visualization includes:

- **Blue shading:** Control Group 95% CI.
- **Orange shading:** Treatment Group 95% CI.
- **Gray/neutral overlap:** Regions where both intervals intersect.

These shaded areas communicate where the *true* conversion rates for each group likely lie.

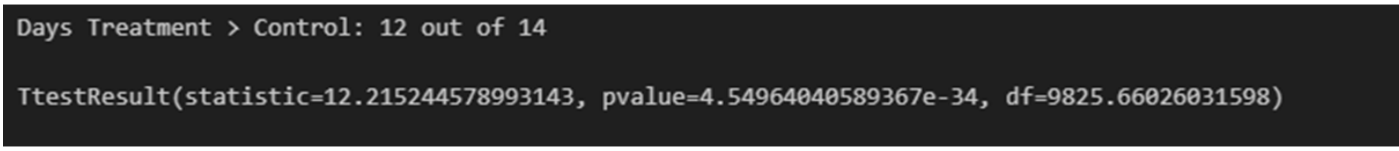
1. The Statistical Objective

When we look at the graph, we are trying to determine if the true conversion rate for the Treatment group is definitely higher than the true conversion rate for the Control group.

- The lines (dots) show the measured conversion rate on that specific day.
- The shading (CI) shows the plausible range for the true conversion rate.

2. Understanding Overlap in A/B Testing

- In statistics, the degree of overlap is a quick visual cue for statistical significance:
- **Large Overlap:** Suggests the difference may be due to random variation, implying weak or no significance.
- **Minimal or No Overlap (the case in my graph):** Indicates a statistically significant difference, strongly suggesting that the Treatment outperforms the Control.



```
Days Treatment > Control: 12 out of 14

TtestResult(statistic=12.215244578993143, pvalue=4.54964040589367e-34, df=9825.66026031598)
```

Across most days, the Treatment Group's confidence band remains distinct and positioned above that of the Control Group. Overlap is small and infrequent, reinforcing the stability and significance of the observed uplift.

Secondary Metric Analysis: Independent Samples T-Test on Clicks

After validating the primary metric (conversion), I conducted a secondary statistical test to determine whether the Treatment Group also generated higher engagement.

Earlier, I observed a difference in the sample means:

$$\text{Treatment Mean Clicks} = 8.07, \qquad \text{Control Mean Clicks} = 6.93.$$

To test whether this difference was statistically significant, I applied a two-sample Welch's t-test from Python‘s scipy.stats (using `equal_var=False`):

- **Null Hypothesis (H_0):** The mean clicks for the Treatment and Control groups are equal.

Technical Interpretation

- **T-Statistic:** 12.215 — This very large value indicates that the difference between the mean clicks of the two groups is substantial in terms of standard error.
- **P-value:** Approximately 0 (i.e., < 0.05), leading to rejection of the null hypothesis.

Conclusion: The Treatment Group's higher average click count is statistically significant. This mirrors the significance observed in the primary conversion metric.

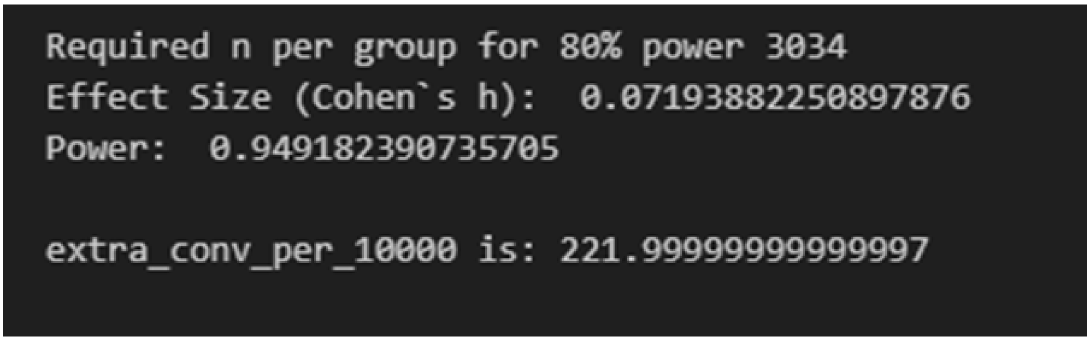
Business Interpretation

This two-metric validation provides high confidence in the feature's overall impact.

- **Improved User Experience (UX):** Higher clicks indicate users feel more comfortable exploring the interface.
- **Increased Engagement:** The new design successfully encourages deeper interaction with the site.

- **Dual Validation:** Both conversion and engagement metrics show statistically significant improvement.

This phase confirms that the new feature not only improves the bottom-line metric (conversion) but also enhances behavioral engagement, strengthening the business case for a full rollout.



```
Required n per group for 80% power 3034
Effect Size (Cohen`s h):  0.07193882250897876
Power:  0.949182390735705

extra_conv_per_10000 is: 221.99999999999997
```

Phase 7: Statistical Power and Future Experiment Planning

This final analytical phase transitions the report from a purely retrospective evaluation (*what happened*) to a forward-looking, strategic assessment (*what we learned and how to plan future experiments*). The objective is to evaluate the statistical efficiency of the A/B test through calculations of Effect Size, Statistical Power, and Required Sample Size. These metrics are fundamental for validating the robustness of the current experiment and guiding evidence-based future testing.

Effect Size: Measuring the Magnitude of Change

The Effect Size is a technical measure of the magnitude of the difference between the new feature and the old one. A larger effect size means the new feature is substantially better

To standardize the magnitude of the conversion lift, I computed Cohen's h using the observed conversion rates:

$$p_{control} = 9.58\%, \qquad p_{treatment} = 11.80\%.$$

The resulting effect size is:

$$\text{Cohen's } h = 0.0719.$$

Interpretation

Cohen's h provides a scale-independent measure of improvement. The value 0.0719 corresponds to a small-to-moderate but statistically meaningful difference between the Treatment and Control groups. While subtle in magnitude, such effects are operationally valuable and highly relevant in high-traffic digital products where percentage-point changes scale significantly.

Business Implication

This metric tells any business exactly how big of a win the new feature is. It quantifies the improvement in a standardized way, allowing for an apples-to-apples comparison against other A/B test results that may have been run.

Statistical Power: Sensitivity of the Experiment

Statistical Power is the probability that my A/B test was sensitive enough to correctly detect the difference (the 2.22% lift) that actually exists. **Ideally, a test should have a power of 80% or higher**

$$\text{Observed Power} = 0.949182.$$

Interpretation

The test achieved a power of approximately 94.9%. With $5,000$ users per group, the sample size was more than adequate to detect the observed effect.

Required Sample Size for Future Tests

To support efficient future experiment planning, I inverted the power calculation to determine the sample size needed to detect an equivalent effect size ($h = 0.0719$) with 80% power.

$$\text{Required Users per Group (80\% Power)} = 3034.$$

Business Implication

This insight directly enhances experimentation efficiency. While the current test used $5,000$ users per group, only $3,034$ were required to reach 80% power. Future tests targeting a similar effect can therefore be shortened substantially, enabling faster decision-making and more rapid product iteration cycles.

Direct Business Value: Incremental Conversions

Finally, I calculated the direct business gain from the 2.22% lift:

$$\text{Extra Conversions per 10,000 Users} = 222.$$

Interpretation

For every $10,000$ visitors exposed to the new feature, the business can expect approximately 222 additional conversions compared to the old feature.

If the platform receives $100,000$ monthly visitors, the feature is projected to generate:

$$222 \times 10 = 2,220 \text{ additional conversions.}$$

This provides a clear, revenue-aligned justification for adopting the Treatment experience.

Conclusion

Phase 7 reinforces that the new feature is a statistically validated, operationally reliable, and financially meaningful improvement. The experiment demonstrates:

- A stable and statistically significant conversion lift of 2.22 percentage points.
- Strong segmentation performance across user types.
- High statistical power (94.9%)
- A reduced required sample size for future tests, enabling faster experimentation.
- A direct business gain of approximately 222 additional conversions per 10,000 users.

Together, these findings provide the company with high confidence to launch the new feature and a clear framework for designing efficient future A/B tests.